

# Character Keypoint-based Homography Estimation in Scanned Documents for Efficient Information Extraction

Kushagra Mahajan, Monika Sharma, Lovekesh Vig TCS Research, New Delhi, India



### Introduction

- Homography estimation for aligning test document images (contains user-filled information) with a template document (unfilled document) for efficient information extraction from documents.
- Propose a novel, robust and memory efficient algorithm for keypoint extraction from scanned or camera-captured document images.
- Uses distinct tips in the individual characters of the document.
- Keypoint correspondences between the template and test documents used to estimate homography. The test document is aligned with the template. Fields in the test document are extracted.
- The extracted text (printed and handwritten) is then read using pre-existing works.

## **Motivation**



- **Question.** How can we get the information added by hundreds of users on forms like insurance forms, bank receipts, job application forms etc. through their images in an automated process?
- □ For facilitating fast retrieval of information, digitization is being used in every aspect of industry.
  - □ Machinery logs in factories
  - □ Contracts in government offices.
  - Scanned or camera-captured document images such as bank receipts, insurance claim forms etc.
- Scanned or camera-captured documents have variations in orientations and illumination, which cause automated information retrieval systems to be prone to errors thereby requiring manual intervention.
- Document alignment reduces errors in automated information extraction and also reduces time and costs for digitization of scanned documents since no manual intervention is involved.

## **Related Work**



- □ Why develop specialized techniques for documents? There are already numerous image alignment techniques in the literature.
  - Direct pixel-based: Lucas Kanade's Optical Flow [1], Lucey et al's work [2]
  - □ Feature-based: SIFT [3], ORB [4] descriptors.
- Some other works that focus explicitly on alignment of documents.
  - Takeda et al [5] uses the centroids of words in the document to compute the features.
  - Block et al [6] exploited structures in the text document like punctuation characters as keypoints for document mosaicing
  - Royer et al [7] explored keypoint selection methods which reduces the number of extracted keypoints for improved document image matching.
- Other relevant works that have been used for image alignment.
  - Rocco et al [8] proposed an end-to-end architecture for learning affine transformations without manual annotation through a siamese architecture.
  - DeTone et al [9] devised a DNN to estimate the displacements between the four corners of the original and perturbed images in a supervised manner, and map it to the corresponding homography matrix.
  - Nguyen et al [10] trains a CNN for unsupervised learning of planar homographies, achieving faster inference and superior performance compared to the supervised counterparts.



Figure 1. Keypoint detection using the SIFT descriptor in straight and rotated images



Figure 2. Keypoint detection using the ORB descriptor in straight and rotated images

## Contributions



- Propose a novel, fast and memory efficient algorithm for robust character based unambiguous keypoint detection, extracted using a standard OCR like Tesseract.
- Demonstrate how existing homography estimation approaches perform poorly when extended to scanned or camera-captured document images. The limitations were analyzed to come up with our methodology.
- □ We show the effectiveness of our proposed approach using information extraction from two real world anonymized datasets comprised of health insurance claim forms.

## Approach





**Figure 3.** Flowchart showing the entire pipeline for information extraction from scanned or camera-captured document images

## Approach (Contd..)



- Take the template and test document images. Resize them to 1600px x 2400px. Then, threshold the images. Thresholding allows us to mitigate the impact of illumination variations.
- Form lists of characters with distinct left, right, top and bottom tip points. For example, characters such as 'A', 'V', 'T', 'Y', 'a', 'v', 'w' etc. have a distinct left tip, and characters like 'V', 'T', 'L', 'Y', '7', 'r' etc. have a distinct right tip.
- Run an OCR engine like Tesseract to detect words in the template and test documents along the their bounding box coordinates. Look for words beginning or ending with the characters in the 4 lists mentioned in the previous step.
- Select such words in the template document. Use neighbourhood information to find accurate corresponding words in the test document.
- □ Find connected components in the bounding boxes of the words. Extract keypoints in the first, last or both the components depending on the characters present at these positions in the corresponding words. The keypoints will be the distinct tip points present in these characters.

## Approach (Contd..)



- Use the corresponding keypoints obtained to estimate the homography. Then, align the test document image with the template document image.
- The user must mark bounding boxes around the fields to be extracted in the template document image. Corresponding fields are automatically extracted in the test document images.
- □ The extracted fields could be printed or handwritten text. Train binary classifier to distinguish printed and handwritten text patches.
- Printed text is recognized using an OCR such as Tesseract in our case.
- □ For handwritten text, we used both the Google Vision API and our own HTR model [13] for recognition. Their accuracies have been compared in tables in the results section later.

### Approach (Contd..)





**Figure 4.** Left, right, top and bottom tips are shown for some of the characters included in the *begCharList*, *endCharList*, *topCharList* and *bottomCharList* respectively.

### **Experimental Dataset**



Evaluated our proposed approach on two real world anonymized document datasets: 

> UPDATE CONTACT INFO ADDRESS CITY: STATE

Template Image

- The first dataset (Insurance1): 15 insurance claim forms and one corresponding empty template form.
- The second dataset (Insurance2): 15 life insurance application forms and one corresponding empty template form.
- Datasets contain variations in illumination, different backgrounds like wooden table and affine transformed document images.

71P

YOU CAN ALSO LIPOATE YOUR CONTACT INTO ON YOUR NATIONARY PET ACCOUNT ACCESS PAGE AT MY PETINSURANCE.COM TREATMENT DATE/S CROM-HOSPITAL/CLINIC NAM

t or the diagnosis has not been determined. Mease do not write "Sa

his Claim Form, I confirm that to the best of my knowledge

16RET3910

tion I have provided is true and correct. I authorize the

nit itemized invoices with your claim form.

MEMBER INFORMATION	No. of pages:	1 MEMBER INFORMATION	
PET NAME: XORO CIT NAME: STA ADDRESS ON FILE: PHO TO THE PHO	TE: ZIP: DNE:	POLICY NUMBER: PET NAME: NAME: Adoress on file:	UPDATE CI ADDRESS: CITY: STATE: PHONE: EMAIL: *TOU CALASO UP PET ACCOUNT ACCO
CLIM DETAILS  RESON FOR VIST, CHCR ALL THAT APPY:	MISSIW PCT HUSPETAL angle lengths, between or influence of highers or work humbers control and list angle for the hypert of them in	CLAIM DETAILS     CLAIM D	In the box below. VETERINARIAN DIAGNOSE? do not fit symptons (for example impre, 1 ingensi. holde a copy of your pre/) therein quarter hough and engine (f) therein therein the symptonic field of the symptonic fiel
Y 6/ 5/29/17 the information	tam fem, i confirm that to the best of my knowledge There provided a true and current. I wathories the st's medical records to hadremick.	MEMBER SIGNATURE and DATE     X     SUBMIT CLAIM FORM and INVOIC     Phase selent your claim by one method     Deplote claim schemisson will store cla     FAX     (714) 989-5600     OR     MAIL     Nationvide Claims De     PO Box 2344     Brea, CA 92822-2344	only. im processing. D No cover sheet necessary 

Test Image

110000000 UCHISALARY MUST SECTION 2. EMPLOYEE COVERAGE BEING REQUESTED IN EVEN \$20,000 UNITS ONLY T TURN IN THIS FORM IF ETHER OF THESE TWO TRANS PERTAINS TO YOU: (1) You are a new ho OMPLETE THIS SECTION should be listed in even \$20,000 DO NOT LIST premium co JFE (Base Life a \$20,000) EED ISSUE & x anual EMENTAL LIFE IN EFFECT is \$500 cost your guerante AND/OR SUPPLEMEN FE BEING REQUESTED TAL COVERAGE DESIRED SECTION 3. AUTHORIZATION (READ BEFORE SIGNING THIS FORM REVIEWER

Test Image

inator's Signature		Date EntitySA	gency Phone Number
Agency Name		Group #	Division #
Agency Mailing Address		City	ZIP Code
Check Employee's Status:	Rehire/Rehire Date	Option Period	See Next Line)
Reason for Midyear Change: HIRE'S ANNUAL SALARY \$		(SALARY MUST BE COMPI	CTED FOR NEW LIRES
	mation Must Re Included for	Now Hires (Life Summary Sheet May I	
		LY PLEASE PRINT NEATLY	
er ID or SSN (NOT Employee ID)	Date of Birth	Ernal Address	Male Ferrale
Name	First Name		Middle Initial
ng Address (New Address?	Yes []No)	City State	ZIP Cod
SECTION 2. EMPLOYE	E COVERAGE BEING	REQUESTED (IN EVEN \$20.00	UNITS ONLY
uaranteed issue amount of Sucole	mental Life Insurance (Guarant	ERTAINS TO YOU: (1) You are a new hin teed issue equals 2 times your annual sa the same amount of life insurance you has	lary at time of employment)
NEW HIRE/REHIRE E	EMPLOYEE	OPTION PERIOD/MIDYEAR C	OVERAGE CHANGE
		COMPLETE THIS	SECTION
COMPLETE THIS S	SECTION STO DTO walks	Amounts should be listed in	
	even \$20,000 units.	Amounts should be listed in DO NOT LIST pres	nium cost.
COMPLETE THIS 3 Amounts should be listed in DO NOT LIST prem LIFE (flasic L/e is \$20,000)	even \$20,000 units.	Amounts should be listed in	nium cost.
COMPLETE THIS 3 Amounts should be listed in DO NOT LIST prem C LIFE (Basic L/e is \$20,000) CANTEED ISSUE (2 × annual	oven \$20,000 units. ilum cost.	Amounts should be listed in DO NOT LIST pres	nium cost.
COMPLETE THIS I Amounts should be listed in	even \$20,000 units. ilum cost. \$ 20,000.00	Amounts should be listed in DO NOT LIST pren BASIC LIFE IN EFFECT	nium cost.
COMPLETE THIS : Amounts should be listed in DO NOT LIST prem C. LIFE passe Line s 320 2001 ANTEED ISSUE (2 x anual a los of enolyment) ILEMENTAL LIFE (maximum his, including your guaranteed is \$2000,000) AL COVERAGE DESIRED	even \$20,000 units. fum cost. \$ 20,000.00 \$ \$ \$	Amounts should be listed in DO NOT LIST per BASIC LIFE IN EFFECT SUPPLEMENTAL LIFE IN EFFECT BASIC ANDICR SUPPLEMENTAL LIFE BEING REQUESTED TOTAL COVERAGE DESIRED	s s
COMPLITE THIS Amounts should be listed in DO NOT LIST prem CC LIFE faced Line statutory at the advective statutory at the advective transmission at the advective statutory at the advective statutory and LCOVERAGE DESIRED SECTION 3. Advectory sequent bit is should EED may sequent bit is should EED may sequent bit is should advective the advective statutory advectory sequent bit is should advectory sequent bit is should advectory sequent bit is should be advective the advectory bit is should be advectory sequent bit is should be advectory advectory bit is should be adve	seen \$20,000 units. Jam cost. \$ 20,000.00 \$ \$ AUTHORIZATION (RE rests and assess given on to an examination by a dynamic to an examination by a dynami	Amounts should be listed in DO NOT LIST perm BASIC LIFE IN EFFECT SUPPLEMENTAL LIFE IN EFFECT BASIC ANDICR SUPPLEMENTAL LIFE BEING REQUESTED	Num cost.  S  S  S  CRML  b habats on which the pre n may be deemed necessary. b) and may constitute group
COMPLETE TWS Amounts should be listed in UFE more than the store UFE more than the store NATEED SSUE or a small and around the store HELENETAL UFE reasoning HELENETAL UFE SITE HELENETAL UFE SITE	even 53000 units. Imm cett. S 20,000.00 S S S AUTHORIZATION (RE ments) and assues ploto on to a subset	Anounts should be tissed in DO ANY LIST gere BASIC LIFE N EFFECT INFORMATING SUPPLEMENTAL VFE BERG REQUESTED TOTAL COVERIAGE DESIRED AD BEFORE SIGNING THIS FG as time to and contain, with they activate supplementation of the signal and the signal supplementation of the signal automation ingle definition of signal automation of the signal automation of the signal automation of the signal automation of the signal automation automation of the signal automation of the DEGAL NEORMATION THE DATE	Hum cost.
COMPLETE THE'S Accounts should be littled in DAVID SUBJECT STATES UP 12 COMPLETE STATES UP 12 COMPLETE STATES STAT	even 53000 units. Imm cett. S 20,000.00 S S S AUTHORIZATION (RE ments) and assues ploto on to a subset	Answirks should be listed to should be listed by Instruction of the second	Hum cost.
COMPLETE THE'S Accounts should be littled in De Nord ULS Trees (TE ance in a stock) ATTEED DISUE of a small trait of anogenetic LINE(TA) (LE (Finishim) LINE(TA) (LE (Finishim	even 53000 units. Imm cett. S 20,000.00 S S S AUTHORIZATION (RE ments) and assues ploto on to a subset	Anounts should be tissed in DO ANY LIST gere BASIC LIFE N EFFECT INFORMATING SUPPLEMENTAL VFE BERG REQUESTED TOTAL COVERIAGE DESIRED AD BEFORE SIGNING THIS FG as time to and contain, with they activate supplementation of the signal and the signal supplementation of the signal automation ingle definition of signal automation of the signal automation of the signal automation of the signal automation of the signal automation automation of the signal automation of the DEGAL NEORMATION THE DATE	Hum cost.

т

#### **Experimental Details**



- Alignment is followed by text field retrieval and classification of the text into printed or handwritten. We train a 5-layer CNN.
  - Patches of printed text obtained from text lines detected by CTPN [11] on a separate dataset, and patches of handwritten text obtained from the IAM dataset [12].
  - □ We obtain a test accuracy of **98.5%** on the combination of the 2 datasets used for experiments.
- Our algorithm is able to handle translations, rotations, and scaling of the test documents.
  - □ For rotations, the system performance is unaffected for rotations upto ±7 degrees in the x-y plane of the image.
  - □ Horizontal and vertical translations range in between ±40% of the document width and height respectively.
  - □ For our datasets, scaling works perfectly when the width and height are varied from 50% to 200% of their original values.

#### **Results**



MEMBER	NFORMATION			No. of pages:
OUCYN			UPDATE CON	TACT INFO write new information below
PET NAM		4	DORESS: ITY:	
NAME:			TATE: HONE:	ZIP:
ADDRESS ON F	N.E.		MAIL:	
			IT ACCOUNT ACCESS P	YOUR CONTACT INFO ON YOUR NAMONIAN LOS AT MY PETINERAWAY, COM
CLAIM DET	AILS OR VISIT, CHECK ALL TH			TREATMENT DATE(S):
<b>WELLNES</b>	S SERVICES	al applit:		FROM
	RY OR ILLNESS DID YOUR		NUNOSE?	TO: HOSPITAL/CUNIC NAME:
		×		
Adaptosis is the imposed, how work that use of Attached" or list INVOICE[5]	retorinarian can help you with the- fiagnosis being treated, your pet d. the services shown on your invisio	e de oot hit symptoms flor degressis, include a ropp e and at the hespital commi t	oxangle Imping, tame I peur pet's treatment pht, or the diagnosis ha	est or infloctore are symptoms of injuries o records and lab results for this visit if there is a not been determined. Picase da not write '
\$	TOTAL	You must sub	nit <u>beined</u> fin	gg with your claim form.
		Do not send o	stimates.	
MEMBER S	IGNATURE and DATE	_		
x		By signing this the information	on I have provided and's medical reco	firm that to the best of my knowled is true and correct. I authorize the rds ta Nationwide.
SUBMIT CL	AIM FORM and INVOK		Per concernent reco	
Please submit Duplicate risk	AIM FORM and INVOK t your claim by one method im submissions will delay cla	only.		INATIONWICK CLAMIS OKPT MOTIS-O
	714) 989-560		et necessary.	
	08			
PC	ationwide Claims De D Box 2344	partment		
Br	ea, CA 92822-2344			
				16RET3910
			(1)	<b>b</b> )
			(1	<b>b</b> )
63.4	elposse 2		(1	<b>o</b> )
	wipo2008_Z		(1)	<b>o</b> )
	el pozze Z		(1)	0)
Ronewood	el pozza _2 		(1	<b>b</b> )
	el pozze Z		(1)	<b>b</b> )
Romonold Alte	nacioni de ci		(1)	0)
8 one asoid .384	nacioni de ci		(1	<b>b</b> )
Romonold Alte	nacioni de ci	20 <sup>2</sup>	(1)	0)
Reasonal Alter Ale	nunganalitzkor 100500709488 20	~ *	(1)	0)
101 101 101 101 101 101 101 101 101 101		20 <sup>3</sup>	(1)	0)
28			(1)	b)
8 алгания 			(1)	<b>b</b> )
28			(1)	b)
5-12- 5-12- 5-12- 12-12-12- 12-12- 12-12-12- 12-12-12- 12-12-12- 12-12-12- 12-12-12- 12-12-12-12- 12-12-12-12- 12-12-12-12-12-12-12-12-12-12-12-12-12-1			(1)	b)
48: 55-12- 55-12- 12: 55-12- 12: 55-12- 12: 55-12- 12: 55-12- 12: 55-12- 12: 12: 12: 12: 12: 12: 12: 12:			(1)	o)
48: 55-12- 55-12 USPET			(1)	<b>D</b> )
Norman Statistics ST-12- ST-12	аналандаг то то то то то то то то то то то то то		(1)	0)
ST-12- ST-12- SS-12- SS-12- VSPLT	аналандаг то то то то то то то то то то то то то		(1)	0)
	Anarona to to to to to to to to to to to to to		(11	0)
	Anarona to to to to to to to to to to to to to		(11	0)
	Anarona to to to to to to to to to to to to to		(11	0)
Norman Statistics ST-12- ST-12	Anarona to to to to to to to to to to to to to		(11	0)
$R_{\rm constant}$ with the second sec	Anarona to to to to to to to to to to to to to		(11	0)
$\tilde{V}_{1,2}$ and $\tilde{V}_{2,2}$ and $\tilde{V}$	аналандог аласанда асанда асан асан		(11	0)
V une control of the	аналандог аласанда асанда асан асан		(11	0)
All Control of the second seco	аналандог аласанда асанда асан асан		(11	0)



(2c)

Practure

Figure 6. Qualitative results for document alignment



Februikannov Phone Num

Bilde Ofen

70.04

Cation Parted

OFTEN DESCENTOTAL CONTINUE CANSE

COMPLETE THIS SECTION Amounts should be listed in even \$25,000 units.

OATE SEGAED

GATE

(2b)

CATE

DO NOT LIST premium

ASIC LIFE IN OFFECT

The mamber must be envolved in Basic Life sourceage i me you panty is the assistant as propagations on the source of th

PLEMENTAL LIFE IN EFFECT

ASIC AND/OR SUPPLEMENTAL PE BEING REQUESTED

TOTAL COVERAGE DESIRED \$

Hew HireMire Date

First Marine

Case Address

#### Results



CHARACTER RECOGNITION ACCURACY FOR FIELDS IN THE FIRST INSURANCE DATASET. COLUMN (A) GIVES THE ACCURACY OF THE PRINTED TEXT, COLUMN (B) SHOWS THE ACCURACY FOR HANDWRITTEN TEXT TESTED ON THE HTR [19], WHILE COLUMN (C) MENTIONS THE ACCURACY OF HANDWRITTEN TEXT USING THE GOOGLE VISION API.

Field	<b>Tesseract</b> (A)	HTR [19](B)	Vision API (C)
Name	98.6%	88%	92.2%
Pet Name	99.2%	89.5%	92.9%
Address	98.3%	80.4%	85.8%
Hospital	98.7%	77.3%	82.5%
Injury	97.1%	78%	82.6%

Table 1

CHARACTER RECOGNITION ACCURACY FOR FIELDS IN THE SECOND INSURANCE DATASET OF APPLICATION FORMS. COLUMN (A) REPORTS THE ACCURACY FOR HANDWRITTEN TEXT TESTED ON THE HTR MODEL GIVEN BY ARINDAM ET AL., WHILE COLUMN (B) GIVES THE ACCURACY OF HANDWRITTEN TEXT USING THE GOOGLE VISION API. THIS DATASET DOES NOT CONTAIN ADDED TEXT IN PRINTED FORM.

Field	HTR Model [19] (A)	<b>Google Vision API (B)</b>
Agency Name	78.7%	83.5%
Agency Address	78.3%	84.6%
First Name	80.1%	84.5%
Last Name	80.7%	86.7%
<b>Applicant Address</b>	78.4%	82.6%
City	81.9%	93.5%
State	83.2%	89.6%

Table 2

#### Conclusion



- ❑ We proposed a character keypoint-based approach for homography estimation using textual information present in the document to address the problem of image alignment, for scanned or camera-captured textual document images.
- □ We cannot use the contemporary machine learning and deep learning algorithms since such documents do not have smooth pixel intensity gradients for warp estimation.
- □ Feature descriptors like SIFT, ORB etc. produce a large number of inconsistent keypoints due to sharp textual edges producing inaccurate keypoint correspondences.
- □ To address these limitations, we create an automated system which takes an empty template document image and the corresponding filled test document, and aligns the test document with the template for extraction and analysis of textual fields.
- The experiments, which were conducted on two real world datasets, support the viability of our approach.

#### References



[1] B. D. Lucas, T. Kanade et al., "An iterative image registration technique with an application to stereo vision," 1981.

[2] S. Lucey, R. Navarathna, A. B. Ashraf, and S. Sridharan, "Fourier lucas-kanade algorithm," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 6, pp. 1383–1396, 2013.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski, "Orb:An efficient alternative to sift or surf." in ICCV, vol. 11, no. 1. Citeseer, 2011, p. 2.

[5] K. Takeda, K. Kise, and M. Iwamura, "Real-time document image retrieval for a 10 million pages database with a memory efficient and stability improved llah," in 2011 International Conference on Document Analysis and Recognition. IEEE, 2011, pp. 1054–1058.

[6] M. Block, M. R. Ortegon, A. Seibert, J. Kretzschmar, and R. Rojas, "Sitt-a simple robust scaleinvariant text feature detector for document mosaicing," Proc. of ICIA2009, pp. 400–403, 2007.

[7] E. Royer, J. Chazalon, M. Rusinol, and F. Bouchara, "Benchmarking keypoint filtering approaches for document image matching," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1. IEEE, 2017, pp. 343–348.

[8] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6148–6157.

[9] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," arXiv preprint arXiv:1606.03798, 2016.

[10] T. Nguyen, S. W. Chen, S. S. Shivakumar, C. J. Taylor, and V. Kumar, "Unsupervised deep homography: A fast and robust homography estimation model," IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 2346–2353, 2018.

[11] Tian, Zhi, et al. "Detecting text in natural image with connectionist text proposal network." *European conference on computer vision*. Springer, Cham, 2016.

[12] U.-V. Marti and H. Bunke, "The iam-database: an english sentence database for offline handwriting recognition," International Journal on Document Analysis and Recognition, vol. 5, no. 1, pp. 39–46, 2002.

[13] Chowdhury, Arindam, and Lovekesh Vig. "An efficient end-to-end neural model for handwritten text recognition." arXiv preprint arXiv:1807.07965 (2018).



### **Thank You!**

Contact us:

TCS Research, New Delhi India

kushagra.mahajan@tcs.com